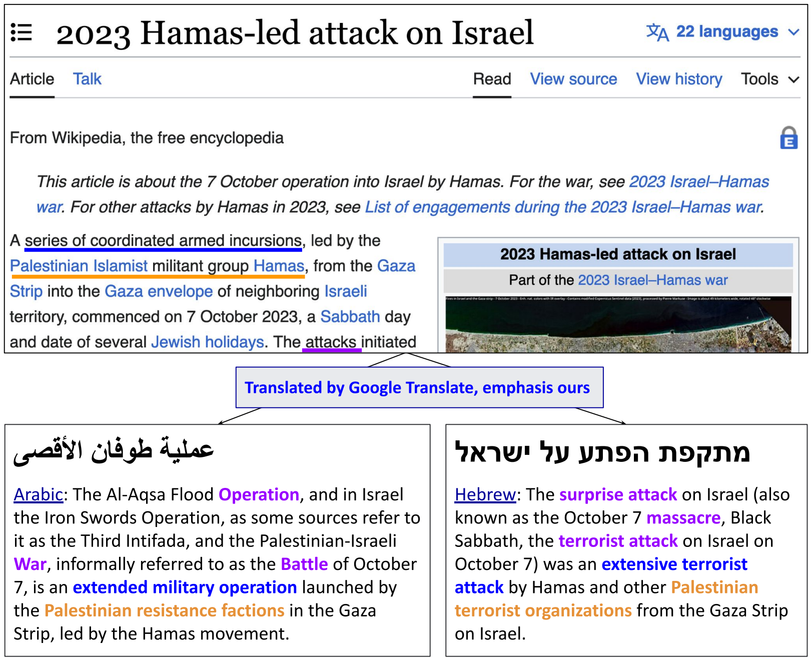
Wikimedia Research Fund 2024 - Stage I proposal

Investigating cross-lingual disparities in representations of conflicts on Wikipedia

Dr Christine de Kock
University of Melbourne



2023 Hamas-led attack on Israel

From Wikipedia, the free encyclopedia

This article is about the 7 October operation into Israel by Hamas. For the war, see 2023 Israel–Hamas war. For other attacks by Hamas in 2023, see List of engagements during the 2023 Israel–Hamas war.

A series of coordinated armed incursions, led by the Palestinian Islamist militant group Hamas, from the Gaza Strip into the Gaza envelope of neighboring Israeli territory, commenced on 7 October 2023, a Sabbath day and date of several Jewish holidays. The attacks initiated

2023 Hamas-led attack on Israel

Part of the 2023 Israel–Hamas war

عملية طوفان الأقصى

Arabic: The Al-Aqsa Flood **Operation**, and in Israel the Iron Swords Operation, as some sources refer to it as the Third Intifada, and the Palestinian-Israeli **War**, informally referred to as the **Battle** of October 7, is an **extended military operation** launched by the **Palestinian resistance factions** in the Gaza Strip, led by the Hamas movement.

מתקפת הפתע על ישראל

Hebrew: The **surprise attack** on Israel (also known as the October 7 **massacre**, Black Sabbath, the **terrorist attack** on Israel on October 7) was an **extensive terrorist attack** by Hamas and other **Palestinian terrorist organizations** from the Gaza Strip on Israel.

Figure 1: The Hebrew, Arabic and English articles on the 2023 Hamas-led attack on Israel.

Abstract

This research proposal seeks to explore disparities in the representations of conflicts across various language versions of Wikipedia. We propose analysing articles across five geographically diverse languages using NLP, with the aim of establishing a robust metric for this phenomenon. The findings will contribute to scholarly discourse on knowledge production and may inform strategies and tools for mitigating biases and unfair representations on Wikipedia.

Introduction

Wikipedia, as one of the largest online repositories of information, plays a crucial role in shaping public perceptions of historical and contemporary conflicts. The Wikipedia guideline on translation discourages automated translations, stating that articles on a given subject in different languages are typically edited independently and need not correspond closely in form, style or content. This can result in divergences such as those shown in Figure 1 for the October 7 attacks, which are described in the

Hebrew version as a terrorist attack and in the Arabic version as a military operation by resistance factions.

Such differences are the natural conclusion of the No original research policy, which states that the community is under no obligation to synthesise all perspectives when there are multiple established views but no authoritative position on a topic. Most readers are likely unaware of the views espoused in other Wikipedia versions, which has implications in terms of bias, fairness and misinformation.

Through this work, we aim to address the following research questions:

1. Which metric(s) best capture divergences in cross-lingual representations on conflicts?
2. Are larger divergences observed between certain pairs of languages or on certain topics?
3. How do these divergences vary over time (during and after a conflict)?

This project addresses the Wikimedia 2030 Strategic Direction priority of Knowledge Equity by breaking down barriers preventing people from accessing knowledge: specifically, the knowledge and awareness of divergent perspectives on a topic across languages.

Dates: July 1, 2024 - June 30, 2025.

Related work

The topic of cross-lingual differences on Wikipedia has been identified in prior work. Callahan and Herring [1] manually coded 60 articles on famous individuals from Poland and the USA. Their results pointed to systematic biases in the English articles of famous people from the USA. Rajcic [2] compared articles about famous individuals for availability in different languages and number of views per article.

Hecht and Gergle [3] introduce an algorithm for identifying analogous sections across multiple languages, finding differences in aspects such as article length, subject matter coverage, and citations.

Though finding cross-lingual divergences, these studies concentrate predominantly on language-agnostic evaluations (such as citations). They do not develop NLP techniques for automatic analysis and quantification of cross-lingual divergence, as we propose here.

Methods

Our exploration will be guided by the data, but we anticipate starting with techniques such as:

- **Span alignment** to identify phrases which discuss related content across two texts.
- **Machine translation or multilingual embeddings** to work across languages.
- **Named entity recognition** of which people, places or things are being discussed.
- **Sentiment analysis** of attitudes towards entities.

We plan to experiment with both available pretrained task-specific models and generalist technologies such as GPT-4. Manual evaluation will be used to judge how closely the automated metrics align with human impressions.

Data

We choose to focus on conflicts because we expect that differences will be especially salient in articles on this topic. We will select articles from the Wikipedia lists of ongoing conflicts, wars, and controversial topics. We have provisionally selected five languages: Hindi, Mandarin, English, Spanish and Afrikaans. This selection is based on *(i)* the availability of first-language speakers within the research group, *(ii)* geographic diversity, and *(iii)* the presence of known historical conflicts.

Expected output

- **Publication:** We expect to publish 2 articles to NLP conferences such as the ACL.
- **Blog:** We believe that readers are not aware of these disparities. A blog post is one

Commented [IO1]: Todo, better transition

Commented [IO2]: At least one of these words is a leftover that should be removed.

possible channel to raise awareness and share results with a wider audience.
- **Tools:** If the analysis indicates large divergences, these findings can be actioned in the form of an editor support tool, an API, and/or banners for readers.

Risks

We foresee the following potential risks:

1. "Disparities", as a phenomenon, is not well-defined. We hope to better characterise it during this project.
2. Our proposal is based on empirical observations, which may not necessarily constitute a broader phenomenon.
3. If it does exist, we may be unable to quantify it accurately using existing NLP tools.

Community impact plan

At a governance level, knowing the extent of cross-lingual divergence is important in informing strategies for promoting fair representation. Systems developed in this work can form the basis of tools for readers and editors. For example, an article-level banner indicating its level of cross-lingual divergence would inform readers' perspective while avoiding the need to arbitrate over the truth. For editors, an interface that highlights an article's divergent parts may be useful.

Evaluation

We plan to use human evaluation to measure the accuracy of our cross-lingual divergence metric(s). Project success will be indicated by the publication of an article in a top NLP venue and an associated blog.

Budget

Our anticipated spending is as follows:

- $30,000: 6 months of RA salary.
- $3,000: Annotation.
- $2,000: Computing.

If there is interest to develop an API/tool, additional funding would need to be secured.

Prior contributions

The PI's extensive prior work on Wikipedia-related projects includes:

- Publishing four first-author papers [4-7] on Wikipedia community dynamics in top, open access, NLP venues.
- Developing Wikipedia resources and sharing them with the research community.
- Interning at the Wikimedia Foundation (2021) and executing a project titled "Effects of collaboration patterns on article quality".

This work will also draw on the sociolinguistics expertise of our advisors. Dr Lea Frermann is an expert in framing and biases in news articles and currently leads a large grant on this topic. Professor Eduard Hovy has extensive experience in computational semantics of language, with a focus on the psycho-social context.

Commented [IO3]: Need to reference the project to explain what DECRA is to non Aussies?

References

[1] E. S. Callahan and S. C. Herring, "Cultural bias in Wikipedia content on famous persons," J. Am. Soc. Inf. Sci., vol. 62, no. 10, pp. 1899–1915, Oct. 2011, doi: 10.1002/asi.21577.

[2] N. Rajcic, "Comparison of Wikipedia articles in different languages," p. 98 pages, 2017, doi: 10.34726/HSS.2017.35937.

[3] B. Hecht and D. Gergle, "The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, in CHI '10. New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 291–300. doi: 10.1145/1753326.1753370.

[4] De Kock, C., & Vlachos, A. (2021). I Beg to Differ: A study of constructive disagreement in online conversations. EACL 2021.

[5] De Kock, C., & Vlachos, A. (2021). Survival text regression for time-to-event prediction in conversations. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.

[6] De Kock, C., & Vlachos, A. (2022). Leveraging Wikipedia article evolution for promotional tone detection. ACL 2022.

[7] De Kock, C., Stafford, T., & Vlachos, A. (2022). How to disagree well: Investigating the dispute tactics used on Wikipedia. EMNLP 2022.